

CITAR PROJECT REPORT

By

Kaustav Saha

Summer Research Exchange Program
Knoesis Research Organization
Dayton, Ohio, USA

May-August 2009

Table of Contents

Section 1 *Motivation*
 1.1 *Current Web Search Issues*.....3

Section 2 *The need for a Semantic Browser and its gradual development*
 2.1 *Need for Semantic Web based Knowledge Exploration*.....6
 2.1.1 *Swanson’s Discovery*
 2.1.2 *Bush’s Memex*
 2.2 *Gradual Development of the Semantic Browser*.....7

Section 3 *Present version of Semantic Browser that exists at the Knoesis Center*
 3.1 *Knowledge Base – background knowledge acquisition*.....8
 3.2 *Spotter-entity mention identification*.....8
 3.3 *Browser-entity/relationship navigation*.....9
 3.4 *Bookmarking-storing trails and sharing results*.....9

Section 4 *Collaboration with the*
CITAR(Center for Interventions, Treatment and Addictions Research) Group
 4.1 *Final Goal*.....12
 4.2 *Problem Statement*.....12

Section 5 *Architecture of the framework developed*
 5.1 *Plan of Action*.....13
 5.2 *User Interface-Ontology Instance Builder*.....13
 5.3 *Backend-Implementation*.....14
 5.3.1 *UML DIAGRAM*.....16
 5.4 *Use of Semantic Browser*.....17

Section 6 *Triple Search and Keyword Search*.....18

Section 7 *Future Approaches*
 7.1 *Machine Learning*.....19

Section 8 *Talks attended at the Knoesis Center,Dayton,Ohio,USA*.....20

***Acknowledgement*.....22**

***Bibliography*.....23**

Section 1 Motivation

Current Web Search Issues

Semantic Web is described as an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation [1]. in a Scientific American article. The evolution of this aspect in Computer Science has heralded a new era in Web Search altogether. In the modern times Search Engines like Yahoo and Google have provided the naïve users with a platform for searching information regarding any topic. But in Computer Science terms the Information content of these search results are far from the User expectations. Unsatisfied with the search results from the query, the user is forced to re-formulate his query, sometimes to a great extent. Still it might be the case that he would not have located the link (website) that he was looking for. This leads to a time consuming and tedious process.

Since most of the web surfers are extremely naïve users finding relevant information is really difficult. The advent of the World Wide Web especially after the year 2000 in a fast pace has created Internet into a huge database of information with poor organization. Most of the searches are query-based. The complete range of behaviours while surfing the Web has been proposed in “taxonomy of WWW user tasks” [2].

Research and study has shown that “expert searchers plan ahead in their searching behaviour based on their knowledge about the Web, while novice searchers hardly plan at all and are rather driven by external representations (what they see on the screen)” [3].

Some Statistical results on the User-Web relationship study elucidates the concerns and discusses these technical issues. Our point of focus is on the naïve users which fall in the category of Web- (lacking in web expertise) and Econo- (lacking in domain knowledge) as the diagrams depict by the light-blue bars.

In Fig-1, this bar graph clearly shows that naïve users do not search topic related website but rather just directly use Search Engines on receiving the particular task.

In Fig-2, this graph shows the actions taken on a Search Engine result page. It shows that naïve users have the highest percentage amongst all the users when it comes to re-formulating the query or altogether creating a new query.

In Fig-3, this graph shows the fact that naïve users have the least percentage in following the link that came from browsing and also the highest percentage in reverting back to the Search Engine.

In Fig-4, this shows that naïve users spend the most time in selecting web-pages.

With E-mail is by far the most common Internet activity, with 90% of all Internet users claiming to be e-mailers [4].

This Statistic puts us a very clear picture that naïve users constitute a huge percentage of Internet Users, therefore the failure of Search Engines in giving pin-pointed search results page (SERP), arises the need for use of Semantic Web to facilitate in searching.

Discussion regarding the need for a semantic browser and its gradual evolution has been done.

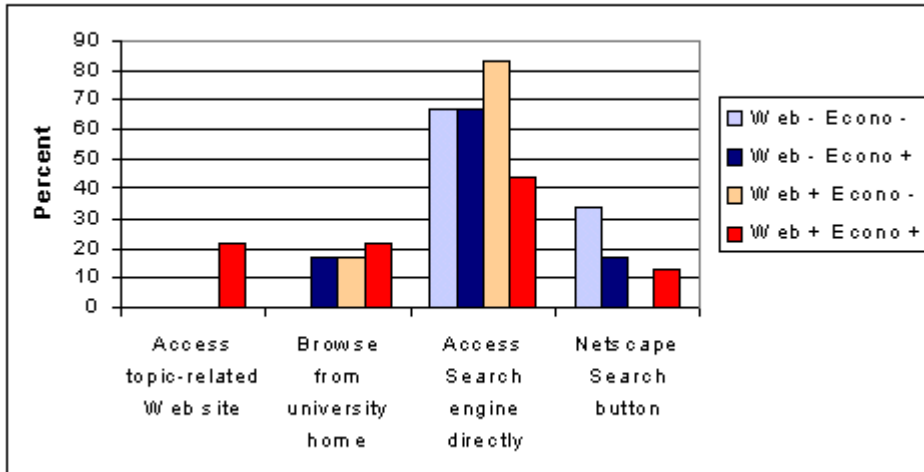


Fig:1-Initial behavior - the first action performed after receiving a task. (Web +/- refers to Web expertise, Econo +/- refers to domain knowledge)

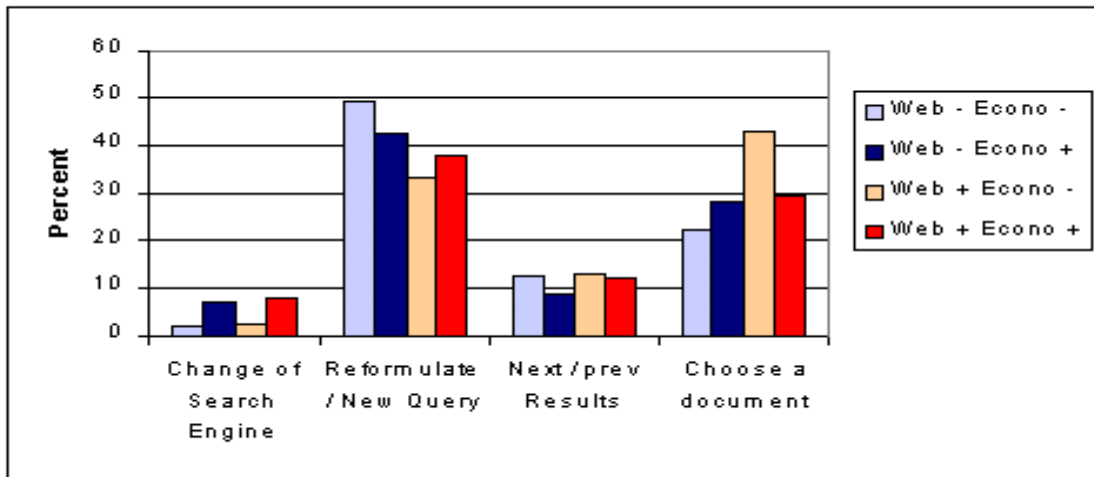


Fig:2-Actions selected on a search engine result page. (Web +/- refers to Web expertise, Econo +/- refers to domain knowledge)

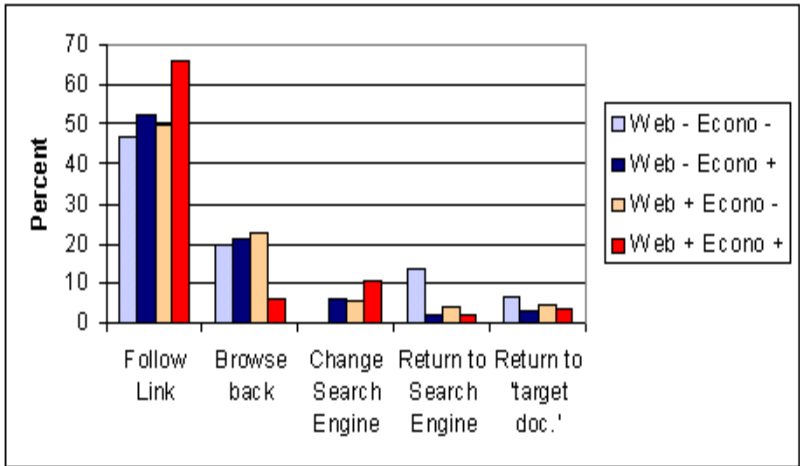


Fig:3-Transitions while Browsing. (Web +/- refers to Web expertise, Econo +/- refers to domain knowledge)

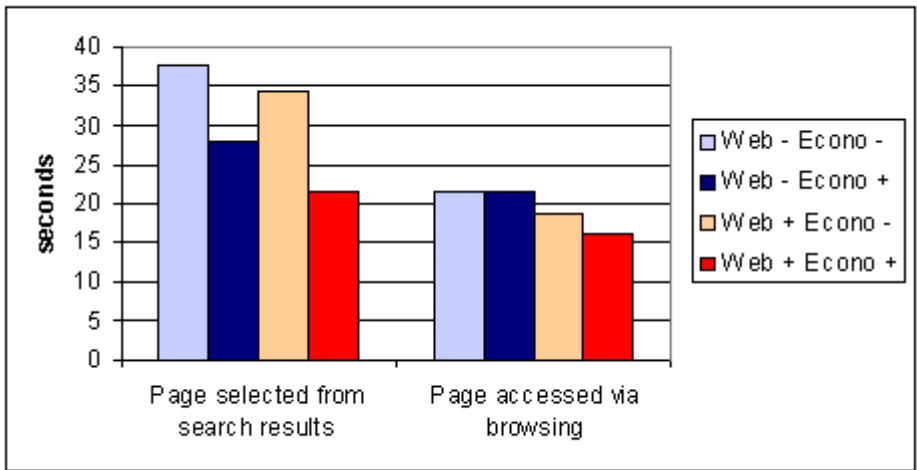


Fig:4-Time spent with Web pages. (Web +/- refers to Web expertise, Econo +/- refers to domain knowledge)

Section 2

The need for a Semantic Browser and its gradual development

2.1 Need for Semantic Web based Knowledge Exploration

The search-and-sift paradigm[5] allows users to go through a large corpus of *Search Engine Results Page(SERP)* returned for a keyword query. The massive amounts of search results is quite over-whelming for the user and he might lose his initial thought process for searching a particular keyword/concept. The user may have come across keywords on a similar context but since the hyperlinks are only pre-established anchors, they make navigation to related contexts extremely difficult. Naive users who constitute maximum percentage of Internet users would start re-formulating their query and some other actions that were explained in the previous section.

The Semantic Browser is a significant step that uses Semantic Web Technologies by keeping track of the user's interaction with the corpus of related web documents. It maintains records of relationships through navigation history of the user. The main idea relies on the technology of finding relationships which are hidden or undiscovered from human cognition through a huge corpus. Linking these relationships and in the process find new relationships or literature, can lead to great scientific breakthrough or discovery as explained in the next sub-section.

2.1.1 Swanson's Discovery

Swanson's discoveries are of great interest in this regard. Swanson started his research from two sets of literature which apparently seemed completely dis-joint-Migraine and Magnesium. From the research literature he found that dietary magnesium supplements leads to specific physiological changes. Those particular physiological changes were essentially pertaining to migraine headaches. From his extensive studies he could formulate associations(11) by which he could infer and in the process discover that "magnesium which is a natural calcium channel blocker inhibits stress which is in turn a symptom observed in some migraine patients"[5].

2.1.2 Bush's Memex

The memex is the name given by Vannevar Bush to the theoretical proto-hypertext computer system he proposed in his 1945 The Atlantic Monthly article [As We May Think](#). The memex has influenced the development of subsequential hypertext and intellect augmenting computer systems.

In Bush's 1945 paper, he describes a memex as an electromechanical device that an individual could use to read a large self-contained research library, and add or follow associative trails of links and notes created by that individual, or recorded by other researchers[6]. He described the concept of "*Trail-blazing*"—the pattern in which the human brain navigates through information space.

He also introduced the concept of "*Associative trails*". The closest analogy with the modern Web browser would be to create a list of bookmarks pointing to articles relevant to a topic, and then to have some mechanism for automatically scrolling through the articles (for example, use [Google](#) to search for a keyword, obtain a list of matches, and then use "open in new tab" in your browser and visit each tab sequentially)[6].

But his vision remained mostly unrealized and extremely limited on the implementation perspective.

The Semantic Browser developed at the Knoesis Research Centre, Dayton, Ohio, USA is a very useful tool for web comprehension based on semantic web technologies. Its gradual development is described in the next section.

2.2 Gradual Development of the Semantic Browser

The Semantic Browser developers have proposed an information exploration mechanism that exploits synergies between search and navigation in the retrieval, selection, organization and comprehension of information on the Web. The following are the key design issues involved.

Keyword search as starting point:Keywords forms the base standard for searching,so it serves as a good initial point.

Enablement of navigation over non-hyperlinked text:Trellis allows named entities identified in text to be treated as anchors for hyperlinks that will point the users to other contexts.The protoype provides a simple dictionary based implementation for entity spotting in text. Trellis architecture allows Named Entity Recognizers (NER) to be easily "plugged-in" according to specific needs.

Customizable navigation:Entities and relationships present in domain models are leveraged as a way to lay down structured information over unstructured text.Through simple mouse commands and pop-up menus the user is able to select entities and explore relationships to find out what is available and to hunt for specific targets. The context of exploration is defined by the domain model and can be exibly controlled by the user. By changing the background knowledge that guides the navigation process, the focus ofthe navigation changes accordingly.

Berrypicking support :The change of user information needs through a series of searches has been shown in [7]. For example,people may shift between information sources as they move from one set of results to another, or they may browse in a source to find a promising new area to explore. In Trellis support for keeping track of the exploration trail (breadcrumbs) has been provided so the user does not lose his/her train of thought.Interface commands such as "pin down" allow the user to select documents of interest before reformulating a query in order to switch context.

Sharing and collaboration:It allows users to organize and share their search result page. In Trellis users have the ability add new results, remove results deemed irrelevant and promote preferred results, generating a list of documents that can be shared with collaborators or saved for later use [5].

Section 3

Present version of Semantic Browser that exists at the Knoesis Center

This section discusses the process to acquire background knowledge and use it for identifying named entities and enabling navigation. The architecture components are in the following sub-sections:

3.1 Knowledge Base – background knowledge acquisition.

The knowledge base (KB) in Trellis stores the domain model that will be used to recognize named entities and guide the navigation. Background knowledge can be acquired in several ways:

- ❖ Importing structured information from existing sources: several knowledge bases exist (e.g. MeSH, UMLS) and can be directly imported into Trellis KB if the knowledge therein can be expressed as triples. Examples of compatible data representation formats include relational databases, RDF, OWL and a number of flat-file and XML data sources.
- ❖ Manually injecting user-generated facts: simple data entry interfaces (e.g. semantic wikis like the Semantic MediaWiki [8]) can be used to gather user-contributed knowledge that can be subsequently imported into Trellis knowledge base.
- ❖ Automatically extracting facts from text: In both supervised and unsupervised ways. General purpose extractors such as the ones described in Ramakrishnan et al. [9] [10] with reasonable success.

Data loading is performed through wrappers that can be provided during deployment as needed. A general-purpose RDF [11] wrapper is provided with the Trellis prototype. Data in the KB is stored in the form of triples "entity 1 - relationship - entity 2". The prototype includes 5232 entities and 16540 triples imported from UMLS[5].

3.2 Spotter-entity mention identification

Named entity recognition (NER), use training data in the form of manually labeled corpora, with tags marking entity mentions. Corpora such as GENIA[12] and BioInfer [13] containing labeled entity had been used. Such tagged corpora are used to collect orthographical [14], contextual [15] and lexical features [15] among others. These features have been shown to perform very well in sequential labeling approaches for identifying specific types of entities like gene names, protein names etc. The prototype implementation provided in Trellis relies on a vocabulary of terms of interest stored at server side. The vocabulary is obtained from thesauri, ontologies, or other sources loaded in our knowledge base. The spotting process uses a prefix tree structure for longest prefix matching. It identifies if a string s composes a named entity in time complexity $O(|s|)$, where $|s|$ is the length of the string s [5].

3.3 Browser-entity/relationship navigation

The Semantic Browser is run by a Java-script application within a common Web browser. The input to the browsing interface is HTML mark-up where named entities are highlighted (enclosed in a tag) by the Spotter module. An example demonstrates the use.

The association between Magnesium and Migraine. The entry-point into Trellis is a keyword search allowing a low cognitive barrier of entry into the system. Entering one or several keywords in the search box. In response to this query, Trellis fetches sentences containing exact and partial keyword matches for each query term, displaying a list of retrieved matches within the workbench. The user selects a named entity. It then creates a pop-up menu that describes the named entity selected along with its id. The user is presented with a description of the selected named entity and the choice to retrieve more documents about that entity, or explore relations to other contexts. The user can choose to view more sentences that contain the original query term by clicking on the hyperlinked query term appearing in the pop-up. In the Relations menu, Trellis displays a list of the entities that are related to the entity by the chosen relationship. On selecting an entity from this menu, the system presents a list of document identifiers. Hovering on a document id, Trellis shows a snippet of the chosen document, along with some metadata to help the user decide whether it is relevant to their information need. Clicking on a document id adds the chosen document to the initial search-engine-result-page (SERP). The inserted document contains a trail from the parent document containing the entity that initiated the query, the initial entity, relationship, related entity and a document identifier of the newly added search result. The user may continue browsing the corpus in this manner building a trellis of possibly intersecting trails. This feature allows the user to keep track of her train of thought while navigating long convoluted paths through the corpus[5].

3.4 Bookmarking-storing trails and sharing results

Document of greater interest found through searching and browsing can be promoted to the top of the SERP, while results deemed insignificant can be removed. Users can bookmark (and unmark) articles to permanently embed them to the workbench. Upon executing another search query, bookmarked items appear before the new search result, demonstrating document persistence during the browsing session. Finally, the user can save the search result, (i.e. top ten documents) for later use. Search results can be downloaded in HTML, contain sentences additional metadata such as the Pubmed URL of the document[5].

Section 4

Collaboration with the CITAR(Center for Interventions, Treatment and Addictions Research) Group

The Center for Interventions, Treatment and Addictions Research (CITAR) is administratively housed within Boonshoft School of Medicine's Department of Community Health, Wright State University. It represents the focal point for substance abuse related services, academic research, and services research. Although the larger purpose of the Center is to advance the production, dissemination and utilization of scientific knowledge and professional technology regarding the epidemiology, consequences, prevention and treatment of substance abuse, its goals are directed at the understanding of substance abuse phenomena and their intervention and management in smaller and mid-sized cities and their surrounding suburban and rural communities[16].

The CITAR Group has been conducting interviews at different places throughout Ohio in places like Dayton, Columbus, Cincinnati etc. Their goal is to understand the social settings in which drug abuse takes place. They want to gather knowledge for specific behaviours which place drug injectors at risk for HIV infection. They have found through their research study that the social contexts in which drug injection occurs, the social roles drug injectors assume, and associated risk behaviours for infection with blood-borne pathogens remain improperly understood[17]. Their study is mainly based on ethnographic research among drug injectors in Ohio.

Ethnographic examination of social roles and the social contexts in which drug injection occurs can complement epidemiological studies and improve AIDS prevention efforts.

“Shooting galleries”, for example are most often described as high-risk settings in which drug users pay a fee (in money or drugs) to enter the premises and inject their drugs as well as a place where previously used syringes are routinely rented by users and then returned to the gallery manager for future re-use[17].

Re-use of needles through needle renting is considered to be very risky for HIV infection. Often these drug users transfer (share) used needles as a means of establishing friendship or a ritualistic form of social bonding.

“Injection Doctors” come to play in this situation. Many of the abusers are new or they lack knowledge, or just have fear of needles or just they are involved in certain kind of relationship. So they are incapable of injecting by themselves as it is difficult for them to locate the vein. Many of the new abusers are heavily influenced by these doctors and their advocating of cheap drugs from streets. The CITAR Group through their extensive research with these “Injection Doctors” have inference that needle re-use is one of the major causes for HIV Infection.

Their study examines initiation to crack injection, methods of preparing crack for injection, the reasons people inject crack, and perceived health consequences[18].

The potency of this “fast-food” form of cocaine in relatively inexpensive doses helped to make crack the perfect commodity for some segments of late twentieth century America. The re-emergence of crack in the early 1980's and its subsequent spread across the nation seems to have been enhanced by the market: a tremendous demand for the stimulating, euphoria-producing and ego-enhancing effects of cocaine supplied at an apparently affordable price. From the marketing perspective, the intensely rewarding, short-duration crack high successfully motivated users to purchase more and more in an attempt to achieve the same effects repeatedly. This income from the sale of crack fuels the supply side, thereby virtually guaranteeing the availability of the drug and the continuation of the epidemic.

Crack smokers experience paranoia. Paranoia is exhibited in strange behaviors such as crawling on the floor, searching for small pieces of crack, or constantly peeking out of the windows looking for police. People resort to extreme measures such as robbing, stealing, deceiving family and friends, or exchanging favors, to obtain more crack.

Some injectors smoked crack to enhance the high achieved through injecting heroin or other opiates, some simply gave up injecting drugs in search of the crack high, and some had difficulty obtaining high-quality heroin or cocaine powder. For others, smoking crack was a way to reduce one's risk of HIV infection through the sharing of injections.

Injectors realized that one could add an acid such as vinegar or lemon juice to crack and prepare a solution for injection. Many of the participants who had a history of heroin or speedball injection were introduced to crack

injection when they had no money or heroin,they were sick and having withdrawal symptoms,and someone else had mixed heroin and crack together.

As a black man said,

“Peer pressure,only reason why I ever done it.Some guy said,’Hey try this.’ You are a drug user,you gonna try.”[18]

Some participants believed the “taste” of lemon juice they derived immediately after injection enhanced their high.Other participants Individuals who preferred vinegar often said that they liked smell of vinegar while preparing the drug solution and enjoyed the “taste” of vinegar once they injected the crack.

Although not mentioned by everyone,people who worked long hours and/or whose jobs or means of making money required increased alertness claimed that injecting crack fulfilled these needs.One white man,for example,injects and smokes crack throughout the night to help himself concentrate,and stay awake while he works.The case study of Shante,a black woman in her 50s,illustrates how the increased alertness she derives from injecting crack cocaine helps her to better perform her “work” [18].

One of the main reasons as proclaimed by these participants was that Crack Injection was less expensive than shooting powder cocaine.The reason for great importance is the one of the greater availability of Crack. Virtually all the participants suggested that one of the major reasons they inject crack is its greater availability compared to powder cocaine.The ease of making “speed-ball” only helps this cause.

The use of illegal drugs is a constantly changing phenomenon that also varies across geographic space. Such changes in the illegal drug-using world have significant public health implications for disease prevention and /or drug-abuse treatment needs,ethnographic monitoring of trends is essential,this is where the research of the CITAR Group comes to importance.

***CITAR**

- Center for Interventions, Treatment and Addictions Research
- Boonshoft School of Medicine
- Wright State University

Personnel

- Dr. Robert Carlson
- Dr. Raminta Daniulaityte
- Mr. Russel Falck

4.1 Final Goal

The research study from the CITAR Group deals with Ethnographic examination of social roles and social contexts in which drug injection occurs can complement epidemiological studies and improve AIDS prevention efforts. They have been interviewing people with pain pills abuse and addiction. On a more high level over-view they would like the Knoesis Research Group to explore their corpus of documents to discover some kind of relationship which would link the following:

“How are pain pills related to HIV ?”

Starting from this research goal, it might be actually possible to discover this relationship between Pain-Pills and HIV through the use of the semantic browser. But this would require a much greater resource of documents and considerable effort in building up the subsequent ontology. From the initial set of interviews that we have worked upon as a base for building up the training-set, the existence of such kind of relationship might seem a bit vague and so to say from a particular perspective not possible computationally. But the beauty of the Semantic Web and other semantic technologies that have been incorporated in the Semantic Browser allows us to link “Magnesium-Migraine” relationship. This is the work that Swanson did through years of research and study of relevant literature and build up “associations” that ultimately lead to this discovery. But the Semantic Browser has been able to discover such a relationship through computational semantics in matter of secs.

So I would infer with a much larger set of documents or corpus, the Semantic Browser would ultimately discover this relationship and solve this research issue.

4.2 Problem Statement

Through a set of meetings conducted with the CITAR Group they have discussed their research issues with us. They have provided us with an initial data-set (10 interviews of drug abusers). Each of the document contains unstructured textual data at great length of the interviews that they have conducted at different places in Ohio. This initial data resource is supposed to act as a good base for developing a training set. At this initial stage in the Project their focus of research was that whether they could computationally discover the following amongst abusers:

“What are the reasons for use of pain pills for drug abuse ?”

From the data-set that they provided us it was possible for us to build a tool

“Ontology Instance Builder”

That would allow us to build the ontology and “plug-in” to the semantic browser to achieve the specific required results. We have been possible to use this CITAR Data-Set to build an Ontology which uses the Semantic Browser to discover the relationship “reason_for_use” for the different drugs (pain pills) mentioned in the interview data source. This work could have been difficult to do entirely manually, use of computational means helps greater ease in finding the relationship and subsequent search or analyzing returned results. Also the user is able to maintain his train of thoughts and look back at previous searches to make out links or discover new kinds of relationship between entities.

Section 5

Architecture of the framework developed

5.1 Plan of Action

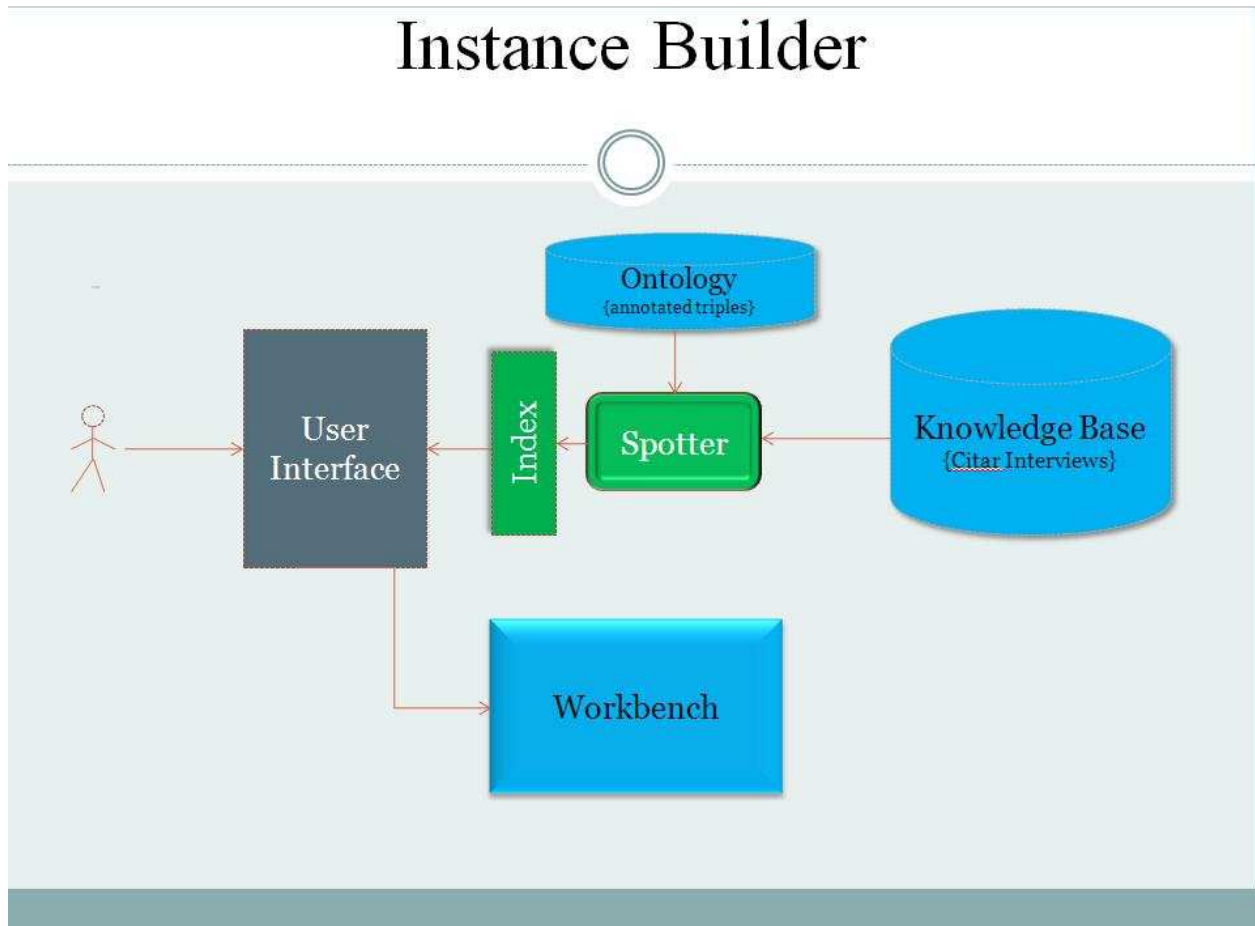
- ❖ Create Data-Set (Subject-Predicate-Object) into an Ontology.
- ❖ Spotter-Identify entities
- ❖ Indexing-Triple Index, Document Entity Index
- ❖ Knowledge Exploration-“Plug-in” to semantic browser.

5.2 User Interface-Ontology Instance Builder

This is the user interface which allows us to create the Ontology by adding Instances (subject-predicate-object) from the Interview Data-Set.

This tool allows users to manually glean through the interviews and create the RDF in the form of subject-predicate-object. It helps the user to locate the entities and the corresponding semantic relationship linked to the object from the un-structured textual data in the interview. The “Search” allows the user to locate a particular keyword (subject) in the ontology and returns back a list of triples from the ontology pertaining to that specific subject. The form contains “Description” which allows the user to provide a comment on the particular subject. It uses “Add Predicate” to provide predicate (semantic relationship), object (option for literal objects is also provided) and Document Id which gives us the document index where this occurs. We can save it in the form by the “Save” button. The “show entities/relationships” gives us a total list of the triples in the ontology that have been entered so far. There is also a ‘delete’ option to remove the particular triple from the ontology. There is also the option “Delete Entity” to remove all triples from the ontology for that particular subject. Then at the end of entering all the triples from the interview set one can put all the data entered into the ontology by the “Export” button.

5.3 Backend-Implementation



cocaine

Search

Subject: cocaine [Delete Entity](#) [Close](#)

Description:

[Add Predicate](#)

Predicate:

Object:

Literal

Document ID:

[\(remove\)](#)

- [cocaine used_by user89](#) [edit](#) [delete](#)
- [cocaine drugs_used_in_combination speed-balling](#) [edit](#) [delete](#)
- [cocaine age_at_first_usage 18](#) [edit](#) [delete](#)
- [cocaine age_at_first_usage 19](#) [edit](#) [delete](#)
- [cocaine candy_dip dip_cigarette_in_cocaine](#) [edit](#) [delete](#)
- [cocaine used_by user94](#) [edit](#) [delete](#)
- [cocaine used_by user82](#) [edit](#) [delete](#)
- [cocaine age_at_first_usage 20](#) [edit](#) [delete](#)
- [cocaine used_by user903](#) [edit](#) [delete](#)
- [cocaine form_used powdered](#) [edit](#) [delete](#)
- [cocaine used_by user99](#) [edit](#) [delete](#)
- [cocaine method_of_consumption snort](#) [edit](#) [delete](#)
- [cocaine streetname white_girl](#) [edit](#) [delete](#)

[\(Show Entities/Relationships\)](#) [Export](#)

5.4 Use of Semantic Browser

The Semantic Browser provides an alternative information exploration mechanism where the user is able to seamlessly “fold” new results into the Search Engine Results Pages (SERP). It can be shown to leverage background knowledge in the form of named entities to provide anchors for navigation, and semantic relationships as interconnections. The system allows the user to navigate from a given entity anchor to a related context in another document. Upon navigating to a new document using such a concept based navigation, it “pulls” the new page into the current SERP for inspection. Users are able to organize (promote, demote, remove) results in the SERP, as well as “pin down” results that should remain visible even after the keywords are changed and a new query is performed. The resulting modified SERP becomes a sharable resource that can serve as a starting point for other users interested in the same topic [5].

Section 6

Triple Search and Keyword Search

Keyword Search: Keyword Search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. The search engines available today provide Keyword Search on top of sets of documents. When a set of keywords is provided by the user, the search engine returns all documents that are associated with these keywords. Typically, two keywords and a document are associated when the keywords are contained in the document and their degree of associativity is often their distance from each other. In addition to documents, a huge amount of information is stored in relational databases, but information discovery on relational databases is not well supported. Through the use of the Semantic Browser we have answered the main question of the CITAR Group at this initial stage of the project

“What is the reason for use of pain pills?”

For example, when the user types in the keyword vicodin, it returns back a list of documents with that keyword. The system allows the user to navigate from a given entity anchor to a related context in another document. Upon navigating to a new document using such a concept based navigation, it “pulls” the new page into the current SERP for inspection. Users are able to organize (promote, demote, remove) results in the SERP, as well as “pin down” results that should remain visible even after the keywords are changed and a new query is performed. The resulting modified SERP becomes a sharable resource that can serve as a starting point for other users interested in the same topic [5]. In this way it can browse through the entire corpus for a particular keyword or its related concept.

Triple Search: It has been possible to incorporate the Triple Search in the Semantic Browser. In Triple Search the user can pull a specific triple from the corpus

{ a particular subject, predicate (relationship) and object }

The present version of the Semantic Browser is based on object-object mapping, so the problem comes when in the auto-fill option provided by the Triple Search component the user starts typing the predicate (relationship) or object, it lists out all the possible options starting with those input characters. But this would create ambiguity in the sense of unavailable triples being searched for.

To alleviate of this technical problem an effort is made to remove this object-object mapping and bring in subject-relationship mapping, relationship-object mapping and subject-object mapping. This would much easily give us the correct options for a subject-relationship or a relationship-object or a subject-object search based on the constraints employed on the respective mappings. Thus only triples will be achieved which are actually present in the corpus and the ambiguity would be completely resolved.

Section 7

Future Approaches

7.1 Machine Learning

A **Markov chain** is a stochastic process with the Markov property. Having the Markov property means that **future states** depend only on the **present state**, and are independent of **past states**. In other words, the description of the present state fully captures all the information that could influence the future evolution of the process. Being a stochastic process means that all state transitions are probabilistic (determined by random chance and thus unpredictable in detail, though likely predictable in its statistical properties).

At each step the system may change its state from the current state to another state (or remain in the same state) according to a probability distribution. The changes of state are called transitions, and the probabilities associated with various state-changes are called transition probabilities. An example of a Markov chain is a random walk on the number line which starts at zero and transitions +1 or -1 with equal probability at each step [19].

A **hidden Markov model (HMM)** is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. An HMM can be considered as the simplest dynamic Bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a *hidden* Markov model, the state is not directly visible, but output dependent on the state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; Even if the model parameters are known exactly, the model is still 'hidden' [20].

With this knowledge as a background I think Machine Learning algorithms like Markov Chain, Hidden Markov Model would help us to take the new data-set to learn from the existing training data-set that has been built from the 10 Pain Pills/Drug abuse Interview provided by the CITAR Group, and build the Ontology Instance automatically by detecting triples from the new data-set.

One Example could be like the words appearing in a specific sequence and applying machine learning algorithms to learn these words as a markov chain, with previous and next words treated as states. When the software system scans through the sentences then it can use this information to detect relationships and entities from the learned data, thus creating automatically creating triples.

This would facilitate the research of the CITAR Group by building a tool which would automatically create triple set and smart enough (through computer semantics) to build up inter-connections through links, thus answering their queries in a much more sophisticated and faster means, rather than using manual text analysis softwares like Nvivo, Folio Views.

Section 8

Talks attended at the Knoesis Center, Dayton, Ohio, USA

Ontologies and data integration in biomedicine May 27, 2009

By
Olivier Bodenreider
Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland – USA

His talk mainly was concentrated on the following issues:

- Sources of information
 - Created by
 - Independent researchers
 - Separate workflows
 - Heterogeneous
 - Scattered
 - “Silos”
- To identify patterns in integrated datasets
 - Hypothesis generation
 - Knowledge discovery
- Bench to Bedside”
- Integration of clinical and research activities and results
- Supported by research programs
 - NIH Roadmap
 - Clinical and Translational Science Awards (CTSA)
- Requires the effective integration and exchange and of information between
 - Basic research
 - Clinical research
- Knowledge management
 - Annotating data and resources
 - Accessing biomedical information
 - Mapping across biomedical ontologies
- Data integration, exchange and semantic interoperability
- Decision support
 - Data selection and aggregation
 - Decision support
 - NLP applications
 - Knowledge discovery
- Warehousing
 - Sources to be integrated are transformed into a common format and converted to a common vocabulary
- Mediation
 - Local schema (of the sources)
 - Global schema (in reference to which the queries are made)
- Linked data
 - Links among data elements
 - Enable navigation by humans

- Role
 - Provide a conceptualization of the domain
 - Help define the schema
 - Information model vs. ontology
 - Provide value sets for data elements
 - Enable standardization and sharing of data
- Examples
 - a. Annotations to the Gene Ontology
 - b. BioWarehouse
 - c. Clinical information systems
- ❖ Conclusion
- Ontologies are enabling resources for data integration
- Standardization works
 - Grass roots effort (GO)
 - Regulatory context (ICD 9-CM)
- Bridging across resources is crucial
 - Ontology integration resources / strategies (UMLS, BioPortal / OBO Foundry)
- Massive amounts of imperfect data integrated with rough methods might still be useful

Acknowledgement

- ❖ *Delroy Cameron, PhD student, Knoesis Center*
- ❖ *Prof. Amit P. Sheth, Director, Knoesis Center*
- ❖ *Dr. Karthik R. Gomadam*
- ❖ *Meena, Topher, Ajith*
- ❖ *Dr. Cartic Ramakhrisnan*
- ❖ *All Lab members- Ashutosh, Raghava, Pramod, Satya, Harshal, Wenbo, Cory, Prateek, Pankesh, Tonya*

BIBLIOGRAPHY

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities, *Scientific Am.* vol. 284, no. 5, May 2001, pp 2837.
- [2] Byrne, M. D., John, B. E., Wehrle, N. S., & Crow, D. C. (1999). The tangled web we wove: A taskonomy of WWW use. In *Human Factors in Computing Systems: Proceedings of CHI 99* (pp. 544-551). Reading, MA: Addison Wesley.
- [3] Christoph Hölscher & Gerhard Strube . Center for Cognitive Science, Institute for Computer Science & Social Research, University of Freiburg, Germany .Web Search Behavior of Internet Experts and Newbies.
<http://www.www9.org/w9cdrom/81/81.html>
- [4] http://www.stanford.edu/group/siqss/Press_Release/press_detail.html
- [5] Pablo N.Mendes, Cartic Ramakrishnan, Delroy Cameron, Amit P. Sheth. Kno.e.sis Center, CSE Department, Wright State University, Dayton OH 45410, USA. Trellis-Knowledge-driven Text Exploration.
- [6] <http://en.wikipedia.org/wiki/Memex>
- [7] M. J. Bates, The design of browsing and berrypicking techniques for the online search interface, *Online Review* 13 (5) (1989) 407-424.
<http://www.gseis.ucla.edu/faculty/bates/berrypicking.html>
- [8] M. Krotzsch, D. Vrandečić, M. Volkel, H. Haller, R. Studer, Semantic wikipedia, *Web Semantics: Science, Services and Agents on the World Wide Web* 5 (4) (2007) 251-261.
<http://dx.doi.org/10.1016/j.websem.2007.09.001>
- [9] C. Ramakrishnan, P. Mendes, S. Wang, A. Sheth, Unsupervised discovery of compound entities for relationship extraction, 2008, pp. 146-155.
http://dx.doi.org/10.1007/978-3-540-87696-0_15
- [10] C. Ramakrishnan, P. N. Mendes, R. A. Gama, G. C. Ferreira, A. P. Sheth, Joint extraction of compound entities and relationships from biomedical literature., *Web Intelligence and Intelligent Agent Technology*, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on, vol. 1, 2008.
<http://dx.doi.org/10.1109/WIIAT.2008.295>
- [11] G. Klyne, J. J. Carroll (eds.), *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C Recommendation, World Wide Web Consortium, 2004.
<http://www.w3.org/TR/rdf-concepts/>
- [12] J. D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus semantically annotated corpus for bio-textmining, *Bioinformatics* 19 Suppl 1.
<http://dx.doi.org/10.1093/bioinformatics/btg1023>
- [13] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Jarvinen, T. Salakoski, Bioinfer: A corpus for information extraction in the biomedical domain, *BMC Bioinformatics* 8 (50).

[14] R. Tsai, C. L. Sung, H. J. Dai, H. C. Hung, T. Y. Sung, W. L. Hsu, Nerbio:using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition, BMC Bioinformatics 7 <http://dx.doi.org/10.1186/1471-2105-7-S5-S11>

[15] P. P. Talukdar, T. Brants, M. Liberman, F. Pereira, A context pattern induction method for named entity extraction, in: Tenth Conference on Computational Natural Language Learning (CoNLL-X), 2006. PDF/cpi_conll2006_camera.pdf

[16] <http://www.med.wright.edu/citar/>

[17] Robert G. Carlson. Shooting Galleries, Dope Houses, and Injection Doctors: Examining the Social Ecology of the HIV Risk Behaviours Among Drug Injectors in Dayton, Ohio. Human Organization, Vol. 59. No. 3, 2000

[18] Robert G. Carlson, Russel S. Falck, and Harvey A. Siegal. Crack Cocaine Injection in the Heartland: An Ethnographic Perspective. Medical Anthropology, Vol. 18, pp. 305-323.

[19] http://en.wikipedia.org/wiki/Markov_chain

[20] http://en.wikipedia.org/wiki/Hidden_Markov_model